

On the Dynamics of IP Address Allocation and Availability of End-Hosts

Oded Argon¹, Anat Bremler-Barr², Osnat Mokryn³, Dvir Schirman¹
Yuval Shavitt¹, and Udi Weinsberg¹

¹ School of Electrical Engineering Tel-Aviv University, Israel

² Computer Science Dept. Interdisciplinary Center, Herzliya, Israel

³ School of Computer Science, Tel Aviv-Yaffo College, Israel

Abstract. The availability of end-hosts and their assigned routable IP addresses has impact on the ability to fight spammers and attackers, and on peer-to-peer application performance. Previous works study the availability of hosts mostly by using either active ping or by studying access to a mail service, both approaches suffer from inherent inaccuracies. We take a different approach by measuring the IP addresses periodically reported by a uniquely identified group of the hosts running the DIMES agent. This fresh approach provides a chance to measure the true availability of end-hosts and the dynamics of their assigned routable IP addresses. Using a two month study of 1804 hosts, we find that over 60% of the hosts have a fixed IP address and 90% median availability, while some of the remaining hosts have more than 30 different IPs. For those that have periodically changing IP addresses, we find that the median average period per AS is roughly 24 hours, with a strong relation between the offline time and the probability of altering IP address.

1 Introduction

Many ISPs dynamically assign IPv4 addresses to hosts, using methods such as Dynamic Host Configuration Protocol (DHCP) and Internet Protocol Control Protocol (IPCP). Both methods enable ISPs to change the routable IP addresses assigned to customers, either when a *lease time* expires (in DHCP) or on customer modem restart (in IPCP).

Understanding the time period during which a host is online and reachable using the same IP address, has implications on various network applications and studies. Specifically, tasks like malicious host identification, network forensic analysis and other blacklisting based approaches require tracking connected hosts over time using their IP addresses [1–6]. The main assumption that lies behind the highly popular blacklists is that a blacklisted machine is identified by its routable IP address, which is completely invalid when facing dynamic IP addresses. Additionally, many peer-to-peer applications, such as file-sharing, voice chats, multi-player games and distributed storage [7], identify hosts using their routable IP addresses, and need it to be stable for prolonged periods in order to provide better service [8].

Capturing these dynamics, however, is a non-trivial task. Previous studies [9, 8, 7] used DNS to gain routable IP addresses of hosts, and then performed long-term probing of these IP addresses. The underlying assumption that the probing is performed on the same host fails if the IP address changes during the time of the probe. Additionally, there is no easy way to tell whether the ping reply came from the end-user machine, its network equipment, or even its ISP equipment (e.g., a gateway). However, these issues are commonly overlooked.

The most similar large scale study of the dynamics of IP addresses was performed by Xie *et al.* [3] who used a month-long Hotmail traces for finding more than 102 million dynamic IP addresses. However, the authors used the Hotmail credentials in order to uniquely identify machines. This is problematic since many users access their emails from many different machines.

A smaller study of a DHCP server was performed by Khadilkar *et al.* [10], who studied the Georgia Tech LAWN (Local Area Walkup and Wireless Network) DHCP server for 5 days. The authors showed a median session time of 75 minutes, which is mainly attributed to the dynamics of people in the campus and wireless networks. We take a broader look, and target various types of ISPs, and the dynamics of availability for mostly stationary hosts.

In this paper, we take a fresh approach to study the dynamics of IP addresses, by leveraging periodical announcements originated from a set of over 1800 end hosts, and further validate our results over a large dataset provided to us by a commercial Internet advertising company. We analyze statistics of the IP address holding time intervals, and host availability. For a better understanding of the interval dynamics, we introduce tools from signal processing that enable us to identify periodic behavior, when exists, and study the periodicity statistics.

Our results show that over a period of 2 months, over 60% of the hosts have a single IP (which we term a fixed IP) and a 90% median online time. Hosts with alternating IPs exhibit a much lower 40% median online time, and some have periodic patterns of IP alternation, with a median period of 24 hours. We also found a strong relation between the length of offline times and the probability that a host will change its IP when coming back online.

2 Measurement Setup

2.1 Dataset

The data used in this paper is obtained from DIMES [11], a community-based Internet measurements system, which utilizes hundreds of software agents installed on user PCs and on PlanetLab servers. In a DIMES installation, a user installs a single agent on each machine. Each agent has a unique ID which is associated with the machine it is installed on.

When an agent is online, it asks for a measurement script from the DIMES central server, and it performs between 2 (set by default) and 4 measurements per minute from this script. Once all measurements are executed, the agents report their results to the server, roughly every 30 to 60 minutes.

There are some exceptions to this normal agent behavior, which introduces “noise” in the dataset. First, the time between consecutive accesses of an agent to

the central server can be different than expected, either due to special measurement scripts of varying sizes, or due to short term network and server failures. We address this by allowing delays of up to 3 hours before considering an agent to be “offline”.

Second, some users duplicate agent settings into additional machines, instead of installing new agents. This results in lower times between consecutive accesses to the server. Furthermore, when the agent is duplicated in machines that use different routable IP addresses, it may appear as if the same agent has constantly altering IPs. We identified one such agent and removed it from the dataset.

Third, since an agent cannot be installed twice on the same machine, some users run several virtual machines and install an agent on each of them. Similarly, some users install agents on a set of hosts that are behind a NAT. In both cases, if standard installation is made, then there will be several different agents that seem to behave exactly the same. If a duplication of a single agent was performed, then it will appear that this agent performs frequent accesses to the server, and we will gain finer sampling. In case of multiple installation, we identify agent clones by comparing their list of IP addresses, and keep only one agent for each set that exhibit identical IPs.

Finally, an agent can be installed on a laptop, which moves between different locations, changing IP addresses, ASes, and possibly countries. Although this may introduce a great deal of noise into the data, we later show that these agents are not common, and there is an easily detectable gap of, at least, a couple of hours between changes in location. The analysis performed is in the granularity of “online” windows, and only when all IP addresses are in the same AS. When an agent exhibits several ASes in the same “online” window, we remove this window from the analysis. We elaborate on this issue in the following section.

Two months of data are used, from June 20th until August 20th in 2010. During this time, the DIMES server was available 97.3% of the time, with a shortest server downtime of 2 hours and the longest of 13.3 hours. We later detail how these affect the analysis and the results.

2.2 Host IP Intervals

For each agent, the DIMES server provides us with an ordered list of the access timestamps, T_i^j and the corresponding routable IP address, IP_i^j , where j is the access (visit) counter. Considering a set of n visits for agent i , we denote its set of accesses as $A_i = \{(T_i^1, IP_i^1) \dots (T_i^n, IP_i^n)\}$.

For a given agent i , we build a list of intervals I_i , which holds a set of consecutive time frames, such that each time frame starts with the first appearance of an IP in A_i and ends in its last consecutive appearance, namely, before a different IP appears. Additionally, we assume that an agent has gone “offline” and thus end the interval if the time difference between consecutive accesses to the server, $1 < j < n$, $\Delta T_i^j = T_i^{j+1} - T_i^j$ is more than a threshold gap G .

In order to determine G , Fig. 1a depicts the average time between consecutive accesses of each agent to the central server. As expected, 90% of the agents have an average inter-access periods of less than 2 hours, and less than 10% are due

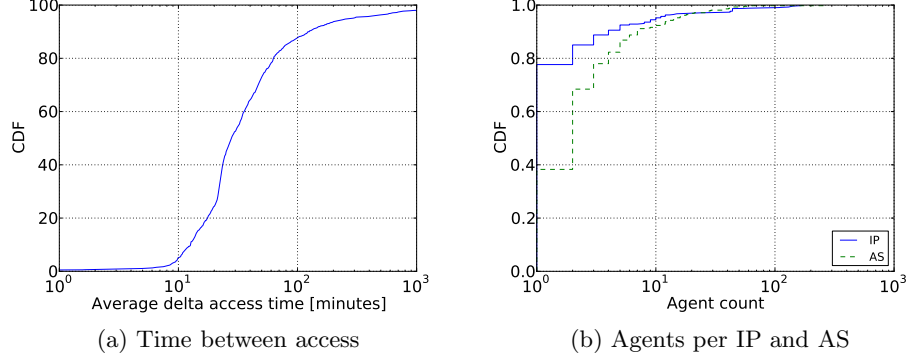


Fig. 1: Cumulative statistics distributions, showing (a) average time between accesses of agents to the central server, and (b) number of agents per IP and AS

to longer scripts, offline periods and server downtimes. Being conservative, we set G to 3 hours, so that we account for the possible further variance in agent access times to the server.

The resulting m intervals for a host i are denoted as the set I , where $I_i = \{(T_i^1, T_i^{e1}, IP_i^1) \dots (T_i^{sm}, T_i^{em}, IP_i^m)\}$. Notice that an interval set may contain the same IP address more than once, either in consecutive intervals (in case there was an offline time between them) or in non-consecutive intervals.

Two consecutive intervals usually have a gap between them, i.e., $\Delta T_i^j > 0$. We define an *online window* as a set of consecutive intervals that have less than G time gap between them, and an *offline window* as the time between any two online windows.

The IP addresses used during gaps in an online window are unknown. Since there is no way to determine which IP address was used and for how long, we split the gap so that its first half is assigned to the interval preceding it, and the second half is assigned to the interval following it. This simple symmetric extrapolation introduces little or no bias to the length of the interval on average, and in any case is bounded by $G/2$, i.e., 1.5 hours.

Finally, there exist agents with intervals that have a zero length, i.e., an IP address that was seen only once within an online window. We found that over 88% of these intervals belong to a single agent, while 95% of the agents have less than 4 zero intervals. By manually examining these zero-length intervals, we found that most of them are due to incorrect identification of the IP address, thus we filtered them out. In the few cases that this filtering method is incorrect, it causes the intervals to appear longer than in reality.

2.3 Host Statistics

Using the two month dataset provides us with roughly 8.6 million samples. Overall, 1804 agents reported 7611 different IPs in 1037 unique address prefixes (AP)

and 432 unique ASes. Agents are spread in 56 countries in all major continents: 41.4% are in the USA, 16% in Western Europe, and 15.9% in Eastern Europe. Some agents are located in remote locations, such as South America (3.4%), Africa (2 agents) and the Far East (4%), but these are mostly PlanetLab servers.

Looking at the types of the ASes in our dataset using [12], we find that 27% of the host are located in academic networks, 3% are in tier-1 ASes, and almost 46% are in tier-2 ASes. The remaining 24% were not resolved, hence they are most likely small regional or corporate ASes, as these are harder to resolve. Therefore, the hosts in the dataset represent a wide range of different commercial and non-commercial networks.

In order to further validate that agents capture a variety of Internet locations, Fig. 1b depicts the cumulative number of agents per IP address and AS. The figure shows that almost 80% of the IP addresses and 40% of the ASes host only 1 agent. The observation that some IPs and ASes host more than 50 agents is mainly the result of a large set of agents behind a NAT, operated by competing groups of users of the DIMES systems, mostly located in Russia and Ukraine. In order to account for this variance, we remove agents that exhibit the same set of IP addresses. Furthermore, when presenting AS-level statistics, we either perform per-AS averaging or normalization prior to the comparison.

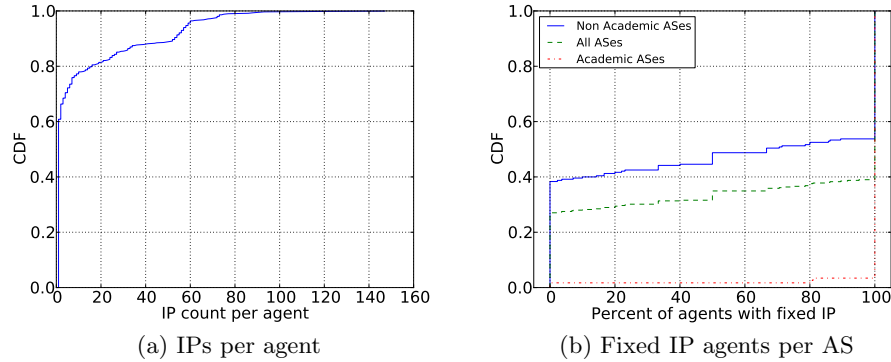


Fig. 2: The number of IPs per agent, and the percentage of fixed agents per AS

2.4 IP Allocation Dynamics

Next, we quantify the percentage of agents that use a fixed IP allocation, i.e., a single IP address throughout the measurement period, and those that exhibit non-fixed IP allocation.

Fig. 2a presents the cumulative distribution of IP addresses per agent, showing that almost 65% of the agents have a single IP address, and less than 5% have 60 different IPs. We further resolved each IP address to its corresponding AS, and found that out of the non-fixed IP agents, almost 85% are changing addresses within a single AS, providing a strong indication that this is not a result of traveling with a laptop, but rather IP alternation performed by their ISPs. In Sec. 3 we further examine the periodical patterns of these agents.

Fig. 2b presents the cumulative distribution of the percentage of fixed IP agents in each AS, showing almost a dichotomy: almost 40% of the non-academic ASes have no fixed IP agents, whereas almost 50% of them have only fixed IP agents. As expected, almost all of the academic ASes have only fixed IP agents.

The high percentage of fixed IP agents is not surprising. Valancius *et al.* [7] showed that most users rarely turn off their modems, hence keep their routable IP address for prolonged times, either by not releasing it or constantly renewing their lease on time. However, the authors also point out that energy-conscious users switch off devices when not in need, a trend that will probably become more common over time, and that will decrease the percentage of fixed IP hosts.

To further validate that indeed there are hosts that exhibit altering IP addresses, and this is not an artifact solely limited to DIMES agents, we used a large dataset provided to us by a commercial company that performs Internet advertising. Using third party cookies, the company uniquely identifies the click-stream of a user in her web-browser. Using a short 4-hour dataset from the last week of Sept. 2010, we analyzed over 16 million unique hosts, performing over 64 million page views. Even in this short time-frame, almost 180k hosts (accounting for 1.1% of the hosts) exhibit more than one IP address, some reaching more than 10 different IP addresses. This further strengthen our observation that IP alternations exists in many of the hosts in today's Internet.

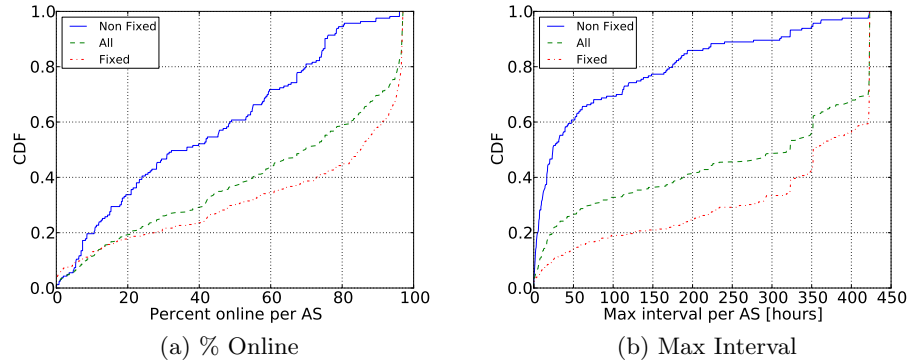


Fig. 3: The percentage of average online time and the maximal interval length, taken over all agents, only fixed agents and only non-fixed agents in each AS

Regardless whether the different IPs are the result of ISP allocation policies, or host mobility, the end result is the same – there is a large portion of hosts that exhibit changing IP addresses. To further understand the availability of hosts in the different networks, we find, for each AS, the percentage of time it stays online. Fig. 3a shows a median online time of 70% for all agents, and a much lower 30% for non-fixed IP agents. No AS reaches 100% online time since the DIMES server was online roughly 97% of this time frame, with a maximal downtime of 13 hours, which is longer than our gap threshold of 3 hours.

These results are lower than those previously reported [7], that measured an average online time of 85% by active probing every 15 minutes. Their higher availability may be explained by the probes reaching gateways, NATs, or Firewalls and not the actual hosts like in our measurements. Alternatively, our results may be lower since the DIMES software itself was stopped. However, this is not common, since due to the low measurement overhead, the agents usually automatically run in the background.

Fig. 3b shows the cumulative distribution of the maximal interval length (i.e., the maximal time that an agent sustained its IP). Notice that the cut-off after 440 hours is because the DIMES server was down for 6 hours roughly 18 days after the beginning of our experiments, hence this value is the maximal possible interval. Although roughly 30% of the ASes have agents that reached the maximal interval, there exists a significant difference between the maximal interval for fixed and non-fixed IP agents, where the latter exhibit a wide range of values, with only 1% that reach the maximal possible interval length.

We further checked the relation between the length of offline times and the probability to have different IPs before and after it. Considering only non-fixed IP agents, the minimum offline time (set to 3 hours) exhibits a probability to change IPs of 0.49, which rapidly increases to above 0.8 after 24 hours. Furthermore, there was no non-fixed IP agent that sustained its IP after an offline time of more than 380 hours.

3 Inferring IP Alternation Period

For each host that has more than a single IP address, we wish to identify whether it has some periodic alternation patterns of IP addresses. Such periodicity can be either attributed to the IP lease times or alternatively to the behavior of the user, e.g., shutdown patterns. Since the data is quite noisy, we use signal processing methods for inferring such periodic patterns, by constructing a time-domain signal of IP alternation, move it to the frequency domain using Discrete Fourier Transform, and extract the dominant frequency.

Constructing the signal is performed at the online window granularity. For each online window, denote by $IP(t)$ the IP address it has during time t , the IP alternation signal $X(n)$ is:

$$X(n) = \begin{cases} X(n-1) & \text{if } IP(nT) = IP((n-1)T) \\ -X(n-1) & \text{if } IP(nT) \neq IP((n-1)T) \end{cases} \quad n = 1..N \quad (1)$$

where $X(0) = 1$. We construct the signal $X(n)$ so that it is discrete, by sampling the IP intervals every minute ($T=1min$). Using one minute sampling is sufficiently fine-grained for capturing any meaningful alternation periods, while keeping low computation overhead.

We then use Discrete Fourier Transform (DFT) over the time-domain signal, for converting it to the frequency domain:

$$Y(k) = \sum_{n=1}^N X(n) \omega_N^{(n-1)(k-1)} \quad , \omega_N = e^{(2\pi i)/N} \quad (2)$$

In case our signal is periodic and *symmetric*, the frequency matching the highest amplitude is most likely to capture its period. Hence, for a given frequency-domain signal, $Y(k)$, the candidate frequency, F , is calculated using:

$$F = \arg \max_k \{|Y(k)|\} \quad (3)$$

The candidate period P is then calculated by $P = 1/(2F)$.

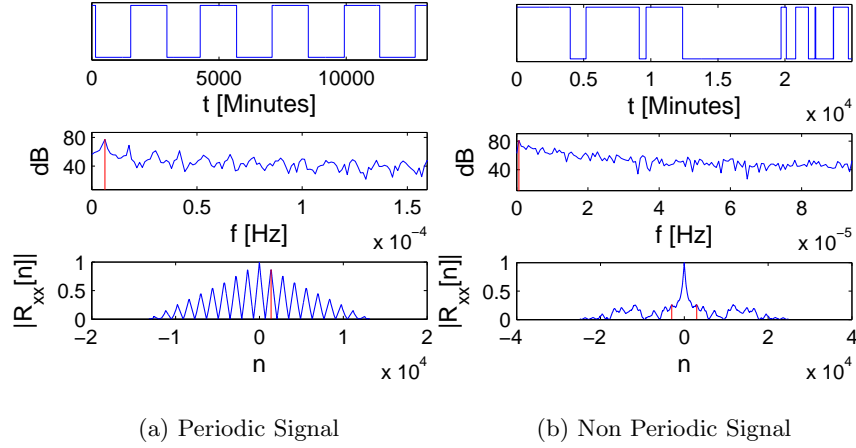


Fig. 4: Samples of periodic and non periodic signals

Several complexities arise when applying DFT to our dataset. First, square signals result in a group of frequencies in the spectral domain (the main frequency and its odd harmonics), lowering the amplitude of the main frequency. Second, DFT is extremely useful for signals with amplitude noise. However, in our analysis, the amplitude is fixed, and the noise resides in the phase of the signal, due to the inability to determine the exact time of IP alternation. Hence, instead of having a constant noise floor, the noise floor of the spectrum depends on the frequency, making it harder to infer the main frequency. Third, extracting P by halving the frequency assumes a duty-cycle of 50%. This seems plausible, as there is not real incentive for having a non-constant IP alternation policy.

The methods we use extract a P for every signal, even if it is not periodic. Therefore, we calculate a confidence value, ξ , that quantifies how close the signal is to really being periodic. We do that by performing auto-correlation of the signal $X(n)$:

$$R_{XX}(n) = \frac{\sum_{m=1}^N X(m)X(m-n)}{N} \quad (4)$$

The normalized auto-correlation is equal to 1 for any signal at the first lag. Periodic signals exhibit high peaks for each multiplication of the period, while for sporadic signals, the auto-correlation peaks are expected to be much lower, and spread across many lags.

Finally, the confidence ξ is calculate as the first (smallest index) peak of the auto-correlation signal, over all lags other than the zero lag:

$$\xi = |R_{XX}(i)|, \text{ s.t. } i = \arg \min_{n>0} \{R_{XX}(n) > R_{XX}(n-1) \& R_{XX}(n) > R_{XX}(n+1)\}$$

Fig. 4 presents two samples of IP alternation signals, their DFTs and R_{XX} . The periodic signal depicted in Fig. 4a has $\xi=0.861$ and a period $P=21.96$ hours, and Fig. 4b depicts a non-periodic signal, with $\xi=0.259$ and a period $P=208.15$ hours. Note the spread of frequencies in the DFT, which is a result of the square signals, and the clear multiplications in the auto-correlation of the periodic signal as opposed to the sporadic signal.

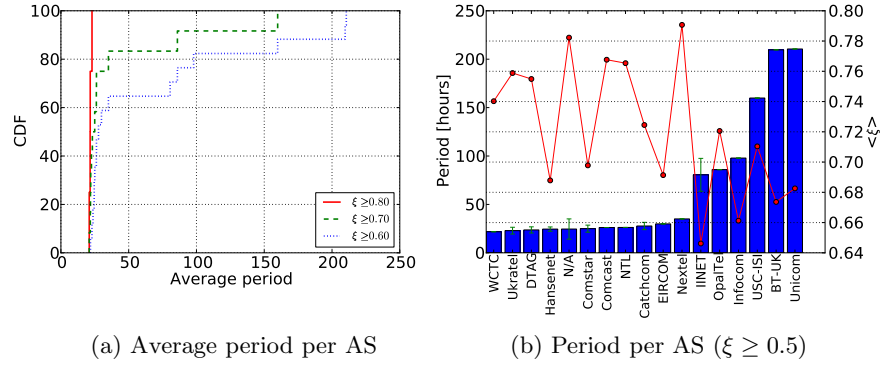


Fig. 5: CDFs of the average period per AS, and AS details using $\xi_{geq0.5}$

Fig. 5 depicts the results of applying the methods on the dataset, while considering only online windows that contain more than 3 intervals, and aggregating to the AS-level. Fig. 5a shows the cumulative distribution of the weighted average period per AS for increasing threshold values for ξ . The average period per AS is taken over all agents contained in that AS, weighted by their average ξ , which is taken over all online windows per agent.

The figure shows that for all three thresholds, over 60% of the ASes have a period that ranges from 20 to 30 hours. As the threshold increases, the average periods become shorter. Surprisingly, the median period is 24 hours, which is similar to the median value seen in Fig. 3b. However, this analysis takes into account the noise, and provides confidence, hence is significantly more reliable.

Fig. 5b provides a breakdown of the ASes for $\xi \geq 0.6$, showing the average period per AS on the left y-axis and the average ξ on the right y-axis. Error bars depict the standard deviation from the average period, which is, as can be seen in the figure, quite small. The figure strengthens the observation that on average, long periods exhibit smaller confidence than short periods. Furthermore, it shows that a rather small number of ASes actually apply periodic policies, and most of them are small regional ISPs or ASes owned by small companies. Using a

threshold of 0.5, 560 agents in 17 ASes were identified as periodic, while for a threshold of 0.8, 185 agents remain in only 4 ASes.

4 Conclusion

This paper presents a measurement study of the dynamics of IP allocation policies in the Internet, which is important for black-lists accuracy and peer-to-peer applications. Using a two-months study, we show that over 60% of the analyzed hosts have a single IP. Using signal processing methods for overcoming inherent noise in the measurements, we found that some of the remaining hosts exhibit periodical patterns of IP alternation, with a median period of 24 hours.

The findings in this paper present a serious issue with the way blacklists are maintained, as a large portion of the blacklisted IP addresses might not be valid after as little as 24 hours. Therefore, any service that relies on a prolonged association of routable IP address and a host must refresh these bindings quite often, or otherwise this matching is most probably wrong.

References

1. Peng, T., Leckie, C., Ramamohanarao, K.: Proactively detecting distributed denial of service attacks using source IP address monitoring. *Networking* (2004)
2. Ramachandran, A., Feamster, N., Vempala, S.: Filtering spam with behavioral blacklisting. In: *The 14th ACM conference on computer and communications security*, ACM (2007) 351
3. Xie, Y., Yu, F., Achan, K., Gillum, E., Goldszmidt, M., Wobber, T.: How dynamic are IP addresses? In: *SIGCOMM*, ACM (2007)
4. Zhuang, L., Dunagan, J., Simon, D., Wang, H., Tygar, J.: Characterizing botnets from email spam records. In: *The 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, USENIX Association (2008) 1–9
5. Ramachandran, A., Feamster, N.: Understanding the Network-Level Behavior of Spammers. In: *SIGCOMM*, ACM (2006)
6. Wilcox, C., Papadopoulos, C., Heidemann, J.: Correlating spam activity with IP address characteristics. In: *Proceedings of the IEEE Global Internet Symposium*, San Diego, CA, USA, IEEE (2010)
7. Valancius, V., Laoutaris, N., Massoulié, L., Diot, C., Rodriguez, P.: Greening the Internet with Nano Data Centers. In: *CoNEXT*, ACM (2009)
8. Cho, K., Fukuda, K., Esaki, H., Kato, A.: The impact and implications of the growth in residential user-to-user traffic. In: *SIGCOMM*, ACM (2006) 207–218
9. Dischinger, M., Haeberlen, A., Gummadi, K.P., Saroiu, S.: Characterizing residential broadband networks. In: *IMC*. (October 2007)
10. Khadilkar, M., Feamster, N., Sanders, M., Clark, R.: Usage-based DHCP lease time optimization. In: *IMC*, ACM (2007)
11. Shavitt, Y., Shir, E.: DIMES: Let the internet measure itself. *ACM SIGCOMM CCR* **35**(5) (2005) 71–74
12. Dimitropoulos, X., Krioukov, D., Riley, G., kc claffy: Revealing the AS taxonomy: The machine learning approach. In: *PAM*. (2006)